NeuroImage 78 (2013) 94–102

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Neural correlates of testing effects in vocabulary learning

Gesa S.E. van den Broek [a,*], Atsuko Takashima [a,b], Eliane Segers [a], Guillén Fernández [b,c], Ludo Verhoeven [a]

[a] Radboud University Nijmegen, Behavioural Science Institute, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands
[b] Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands
[c] Radboud University Nijmegen Medical Centre, Department of Cognitive Neuroscience, P.O. Box 9101, 6500 HB, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Accepted 27 March 2013
Available online 8 April 2013

Keywords:
Testing effect
Retrieval-enhanced learning
Vocabulary development
Subsequent memory
fMRI

ABSTRACT

Tests that require memory retrieval strongly improve long-term retention in comparison to continued studying. For example, once learners know the translation of a word, *restudy* practice, during which they see the word and its translation again, is less effective than *testing* practice, during which they see only the word and retrieve the translation from memory. In the present functional magnetic resonance imaging (fMRI) study, we investigated the neuro-cognitive mechanisms underlying this striking *testing effect*. Twenty-six young adults without prior knowledge of Swahili learned the translation of 100 Swahili words and then further practiced the words in an fMRI scanner by restudying or by testing. Recall of the translations on a final memory test after one week was significantly better and faster for tested words than for restudied words. Brain regions that were more active during testing than during restudying included the left inferior frontal gyrus, ventral striatum, and midbrain areas. Increased activity in the left inferior parietal and left middle temporal areas during testing but not during restudying predicted better recall on the final memory test. Together, results suggest that testing may be more beneficial than restudying due to processes related to targeted semantic elaboration and selective strengthening of associations between retrieval cues and relevant responses, and may involve increased effortful cognitive control and modulations of memory through striatal motivation and reward circuits.

Neural correlates of testing effects in vocabulary learning

Tests that require memory retrieval improve long-term retention more than continued studying (Roediger and Karpicke, 2006b). For example, once learners know the translation of a word, *restudy* practice, during which they see the word and translation again, is less effective than *testing* practice, during which they see only the word and retrieve the translation from memory (Karpicke and Roediger, 2008). This *testing effect* has received much attention from behavioral studies, but its neural correlates are still largely unknown (Roediger and Butler, 2011).

To the best of our knowledge, only two fMRI studies have, so far, explicitly investigated testing effects. Eriksson et al. (2011) scanned participants during a final recall test following prior testing practice, and interpreted correlations between anterior cingulate activation and the amount of prior testing in terms of enhanced memory consolidation. Hashimoto et al. (2011) investigated brain activity related to repeated testing and showed both repetition enhancement and attenuation at the final recall. Both of these studies documented facilitated retrieval processes *after* prior testing. In the present study, we took a different approach and investigated the testing practice phase itself. We directly compared the brain activity related to testing and restudying in order to gain insight into the neuro-cognitive mechanisms by which testing improves memory more than restudying.

Most explanations of testing effects assume that testing improves memory more than restudying because it involves more effortful semantic processing (Roediger and Karpicke, 2006b). More specifically, testing is thought to enhance cognitive effort (e.g., Pyc and Rawson, 2009), which is defined somewhat vaguely as an index of the amount of goal-directed, non-automatic processing (Roediger and Butler, 2011). In this context, testing has also been said to constitute a *desirable difficulty* during learning because it increases beneficial deep semantic processing (Bjork and Bjork, 1992). This could lead to a strengthening of the association between retrieval cues and target information and an improved efficiency of search processes during later recall (e.g., Karpicke and Smith, 2012; Karpicke and Zaromb, 2010), such that irrelevant associations are suppressed and target information comes to mind earlier in response to retrieval cues (Thomas and McDaniel, 2013). Alternatively, testing could improve memory because searching for the correct answer during memory retrievals extends semantic networks around the target information with additional associations, thereby increasing the number of available retrieval cues that can lead to later recall (Carpenter, 2009).

Although these explanations of testing effects are rather abstract, some predictions about possible neural substrates can be derived. First, the inferior frontal gyrus (IFG) has consistently been related to controlled, effortful processing during memory retrieval (Race et al., 2009). More specifically, IFG is thought to maintain retrieval plans

* Corresponding author. Fax: +31 24 3616211.
E-mail address: g.vandenbroek@pwo.ru.nl (G.S.E. van den Broek).

1053-8119
http://dx.doi.org/10.1016/j.neuroimage.2013.03.071

to favor the activation of relevant information, and to be involved in the selection among competing representations (Badre and Wagner, 2007). Furthermore, IFG activity has been related to semantic processing (Gabrieli et al., 1996; Wagner et al., 1998), during which frontal control processes are thought to act on semantic representations stored in more posterior regions of the brain (Whitney et al., 2011). Although semantic representations are probably distributed across multiple brain areas, a recent meta-analysis of 120 studies suggested that the middle temporal gyrus (MTG) and the inferior parietal lobe (IPL) could function as association areas that integrate different aspects of semantic concepts (Binder et al., 2009). More specifically, MTG and IPL seem to mediate the storage and retrieval of word meaning and the integration of information into larger units for semantic processing (Lau et al., 2008). Therefore, it is likely that the coordinated activity of IFG, MTG, and IPL is involved in testing if effortful, elaborate semantic processing enhances the memory trace.

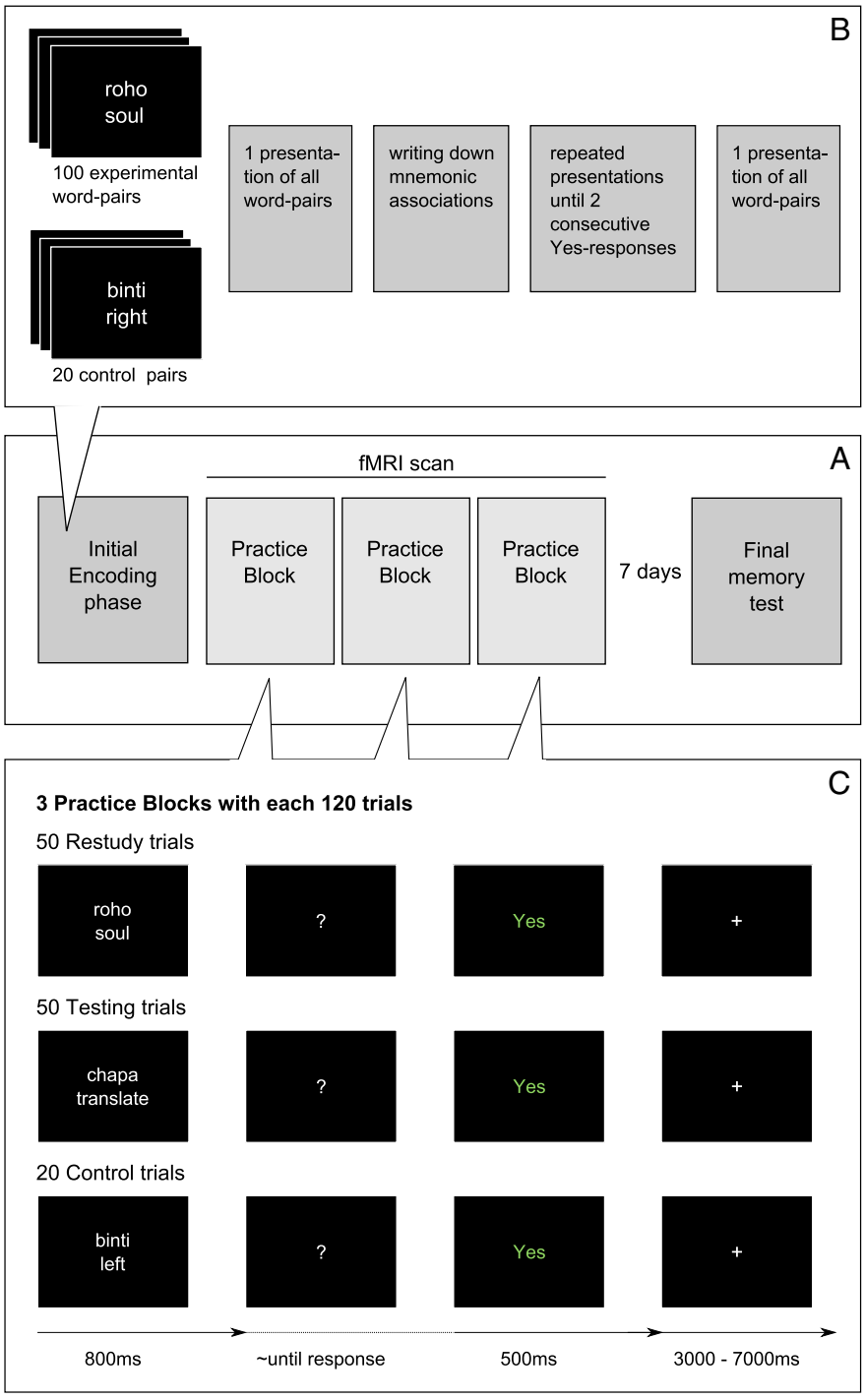

Fig. 1. Experimental procedure. A. Overview of the complete experiment that consisted of an extensive initial encoding phase before scanning, testing and restudy practice in the MR scanner, and a memory test one week later. B. Overview of the four initial encoding tasks with which the participants studied 100 experimental words and 20 control words. C. Overview of the practice trials in the fMRI scanner. This phase contained the critical experimental manipulation: 50 word-pairs were presented in a testing condition with retrieval opportunity, and 50 word-pairs were presented in a restudy condition. In the response phase of both testing and restudy trials, participants pressed a button to indicate whether they thought that they knew the translation of the Swahili word. The response (Yes or No) was displayed for 500 ms. Note that all non-Swahili words were presented in the participants' native language Dutch during the experiment.

To test our predictions about neural correlates of testing effects, we collected fMRI data while Dutch participants practiced previously encoded Swahili-Dutch word-pairs by looking at the whole pair (*restudying*), and while retrieving the translation from memory upon seeing only the Swahili word (*testing*) (Fig. 1C). Based on earlier studies (e.g., Roediger and Karpicke, 2006a), we expected that testing would lead to better recall than restudying on a later memory test. With respect to brain activity, we derived two hypotheses from the idea that testing increases semantic elaborations and effortful cognitive control: First, we expected higher activity in IFG, IPL, and MTG during testing than during restudying. Second, we expected that activity in these areas during testing and perhaps also during restudying would predict later recall.

Materials and methods

Participants

Twenty-six female first-year university students (M_{age} = 19.5 years, SD_{age} = 1.9) participated in the experiment for course credits. The native language of all participants was Dutch and they had no prior knowledge of Swahili. All participants reported that they were right-handed, had normal or corrected-to-normal vision, no neurological or psychiatric history and no language-impairments. The data of 22 participants were included in the analyses; the other four participants were excluded because they had too few trials in specific conditions of interest (i.e., less than ten remembered or less than ten forgotten words). To increase motivation, there was a small financial reward (10 Euro) for the 10% of participants who performed best.

Stimuli

The stimuli were 100 Swahili nouns with their Dutch translation, and 20 control words (also Swahili nouns) of which no translation was given, but which were randomly paired with the Dutch word for "left" or "right". All Swahili words were pronounceable for Dutch native speakers, e.g. "kiti" (chair), "panya" (mouse).

Procedure

The experiment consisted of two sessions, which were both conducted at the same laboratory. Session 1 began with an extensive initial encoding phase, followed by testing and restudy practice in the MR scanner. There was a delay of about 15 min between the initial encoding phase and the practice phase in the scanner, due to preparation of the participants for scanning. Session 2 was conducted one week later, and contained the final memory test (see Fig. 1A).

Initial encoding

The purpose of the initial encoding phase was to let the participants learn the translations of the 100 Swahili words. For this purpose, they studied the Swahili–Dutch word-pairs at the computer with four different tasks (Fig. 1B). Throughout these tasks, the Swahili words were presented simultaneously with their translation to minimize opportunities for retrieval during initial encoding. First, the participants saw all word-pairs once for 8 s each and were instructed to think of an association to remember the words. Second, they typed in a short description of each association when cued with the complete word-pairs. Third, the participants practiced with an adaptive computer program that presented the complete word-pairs, one at a time. After each presentation, the participants were asked to make a judgment of learning by pressing a button for either "Yes, I already know the translation" or "No, I don't know the translation yet". Presentations of each word-pair continued until the participants had responded with "Yes" in two consecutive encoding rounds. The number of rounds necessary to learn each word was then used to assign the words to the experimental conditions in such a way that the mean number of rounds during initial encoding was equal for the 50 restudied words and the 50 tested words for each participant. The control word-pairs (Swahili words paired with the word "left" or "right") were presented during the first two encoding rounds and the participants responded by pressing the indicated (left or right) button. The participants were told that they did not have to remember the control words. Fourth, at the end of the encoding phase, all word-pairs were presented one more time and participants again pressed a button to make learning judgments. In total, the initial encoding phase took about 1 h and 15 min, with variations depending on the number of rounds that the participants required to learn each word.

Testing and restudy practice in the fMRI scanner

The critical experimental manipulation took place in the fMRI scanner, where the participants practiced 50 words in a testing condition and the other 50 words in a restudy condition. The difference between the conditions was that the complete word-pair was visible on the screen in the *restudy* condition, whereas only the Swahili word was visible in the *testing* condition, together with the word "translate". In both conditions, the participants responded by pressing a button with their left hand to indicate whether they knew the translation (see Fig. 1C for details on the timing of the trials). There was no other overt response. Participants were instructed to do their best to further improve their memory for the presented words during scanning and to devote enough attention to each word to make a good judgment of whether they knew the translation of the word or not. The 20 *control* words were randomly paired with the word "left" or "right" in every practice block, and the participants responded with the left or right button. The participants completed three practice blocks in the fMRI scanner, in each of which they saw all 120 word-pairs once in the assigned condition, in a randomized order. Each practice block took approximately 17 min.

Final memory test

Seven days after scanning, the participants took a computerized test, during which they saw the trained Swahili words in a randomized order (one word at a time) and were instructed to type in the Dutch translation. There was no time pressure during responding.

Behavioral data analysis

Responses on the final test were categorized as either correct or incorrect. In addition, response times were obtained by covertly recording how long it took the participants to fill in the translation and click on a button to proceed to the next word, after the Swahili word had appeared on the screen. Only response times for correct responses were analyzed.

MRI data acquisition

A 3 T MR scanner (Magnetom TIM TRIO, Siemens Medical Systems, Erlangen, Germany) was used to acquire T2*-weighted images of the whole brain with an echo-planar imaging (EPI) sequence (35 slices, slice thickness: 3.0 mm, slice gap: 0.3 mm, ascending slice acquisition, repetition time (TR) = 2.22 s, echo time (TE) = 30 ms, flip angle = 80°, matrix size = 64 × 64, field of view: 212 mm). In addition, a structural T1-weighted image was obtained using a magnetization-prepared, rapid-acquisition gradient echo sequence (192 slices, slice thickness: 1.0 mm, TR = 2300 ms, TE = 3.03 ms, flip angle = 8°, matrix = 256 × 256, field of view: 256 mm).

MRI data analysis

Image preprocessing and statistical analyses were performed with SPM8 (*Statistical Parametric Mapping*; Wellcome Department

of Cognitive Neurology, London, UK; www.fil.ion.ucl.ac.uk) implemented in Matlab 7.11 (MathWorks, Natick, MA).

Preprocessing

The first five volumes of each participant's functional EPI data were discarded to allow for T1 equilibration. The EPI images were realigned to the participant mean EPI image, which was co-registered to the corresponding structural image. Both functional and structural scans were spatially normalized to a common Montreal Neurological Institute (MNI) reference brain as defined by the SPM8 T1.nii template (resampled at voxel size 2 × 2 × 2 mm), as well as spatially filtered by convolving the functional images with an isotropic three-dimensional (3D) Gaussian kernel (8 mm full width at half maximum). Slow signal drifts were removed with a high-pass filter with a cutoff period of 128 s.

Statistical analyses

As a first step, the data were analyzed separately for each participant for each of the three practice blocks. Trials were categorized based on the practice condition (testing, restudy, control) and the result at the final test (LR, LF): Later remembered testing trials (LR_T), later forgotten testing trials (LF_T), later remembered restudy trials (LR_{RS}), later forgotten restudy trials (LF_{RS}), and control trials (C). Only practice trials in which the participants responded with "Yes, I know the translation" were used; trials with the answer "No, I don't remember" were modeled as trials of no interest in a separate sixth category. Neural activations corresponding to the six categories were modeled by separate stick functions, which were time-locked to the presentation of the word-pairs and convolved with a canonical hemodynamic response function and its temporal derivative provided by SPM8, to yield twelve regressors in a general linear model of the BOLD response. The design matrix also included six head motion regressors (three translations and three rotations determined from the realignment step). Parameter estimates were calculated and summarized in contrast images against the control trials: $LR_T - C$; $LF_T - C$; $LR_{RS} - C$; and $LF_{RS} - C$. In the second step, these single-subject contrast images were included in a group-level ANOVA with the factors Block (1, 2, 3), Practice Condition (Testing, Restudy) and Memory (LR, LF), in which the participants were treated as random factors. For the statistical analyses, we used an uncorrected threshold of $p < .001$ at voxel-level, and applied a threshold of $p < .05$ (family wise error corrected) at the cluster-level (cf., for example, Hayasaka and Nichols, 2003).

Results

Behavioral results

Initial encoding

Prior to scanning, the participants studied all 100 experimental words and translations to the same criterion (see Materials and methods section for details). The words were then, for each participant, assigned to the two practice conditions in such a way that the average number of presentations during encoding was identical for the 50 tested and the 50 restudied words (across participants $M_{testing} = 3.3$ ($SD = 2.11$), $M_{restudy} = 3.3$ ($SD = 2.14$)).

Practice phase in the scanner

During testing- and restudy practice in the scanner, participants responded with "Yes, I know the translation" to on average 91.1 of the 100 experimental words (the rest of the words were modeled as trials of no interest in the analysis of fMRI data, as described in the Materials and methods section).

Translation performance after one week (Fig. 2)

At the final memory test seven days after practice, participants recalled more translations of the tested words than of the restudied words, $t(21) = 7.436$, $p < .001$, $d = 1.62$. The average performance difference between the two conditions was 8.2%. At the same time, participants were on average 596 ms faster to (correctly) fill in translations of tested words than of restudied words, $t(21) = 3.257$, $p = .004$, $d = 0.71$. When only those words were taken into account to which participants responded "Yes, I know the translation" during practice, the performance difference on the final test increased to 12.3%, $t(21) = 8.682$, $p < .001$, $d = 1.89$, whereas response time differences remained approximately the same ($M_{T\text{-}RS} = 593$ ms). In sum, behavioral testing effects were large and were found both in terms of the amount of information that was remembered and in terms of response times (Fig. 2).

Neuroimaging results

Testing versus restudy

To determine which regions were differentially activated during testing and restudying, the two conditions were compared in a factorial design (see Materials and methods section for details). As shown in Table 1, when trials were combined across the three practice blocks and across levels of subsequent memory, testing engaged a large set of brain areas in comparison to restudying (see also Fig. 3A). This included bilateral anterior and mid-IFG in pars orbitalis (~BA 47; local maximum (hereafter abbreviated) [−30; 24; 0]) and pars triangularis (~BA 45; [−40; 24; 24]), and the left posterior IFG in pars opercularis (~BA 44; [−42; 4; 34]). Other regions that were more engaged during testing than during restudying included the bilateral ventral striatum [12;10; −2] and midbrain areas [8; −20; −12], left supplementary motor areas [−6;18;50], left middle occipital gyrus [−26; −72; 42] and bilateral lingual gyrus [−8; −82;10].

The results for the reversed comparison of restudy over testing trials are reported in Table 2. The right IPL [50; −70; 28] and left IPL [−54; −66;42]; the right MTG [64; −16; −14]; the right middle cingulate gyrus [8; −50;40], right middle frontal gyrus [28; 34; 48] and left middle orbital gyrus [−8; 58; 4] were more active during restudying than during testing.

Practice effects

The results on the final memory test seven days after practice were used to categorize the practice trials into practice of later-remembered (LR) and later-forgotten (LF) words, which were then compared to each other to find areas in which activity predicted later memory. Note that we refer to this contrast as "practice effect" to distinguish it from classic subsequent memory effects (e.g., Kim, 2011), which are based on data obtained during a single encoding opportunity and not during additional practice, as in the present study.

For the restudy items, the LR–LF comparison revealed activity in the bilateral rectal gyrus extending to left superior orbital gyrus [−2;40; −2]. For the testing items, a large set of brain areas was predictive of later memory, including the superior medial and superior frontal gyrus [−12; 56; 8], the left middle cingulate cortex and left precuneus [−6; −56;22], the left and right middle temporal gyrus ([−46; −56;28] and [56; −14; −16]), and the left and right inferior parietal lobe ([−54; −58;30] and [52; −50;40]). The reversed contrast, LF–LR, showed no significant clusters for the restudy items and showed activity in the occipital lobe [−10;78,10] and in the supplementary motor area [−6;8;56] for the testing items.

Differences in practice effects between testing and restudy trials. Practice effects were visible in different areas for the testing and the restudy trials. To test in which brain areas this difference was significant, we calculated interaction effects between practice condition and later memory. The interaction effect showed areas in the supramarginal and angular gyrus in the left IPL ([−56; −46; 44] and [−54; −60; 44]) and the left MTG [−64; −46; −6] (statistics in Table 3, activation map in Fig. 3B)

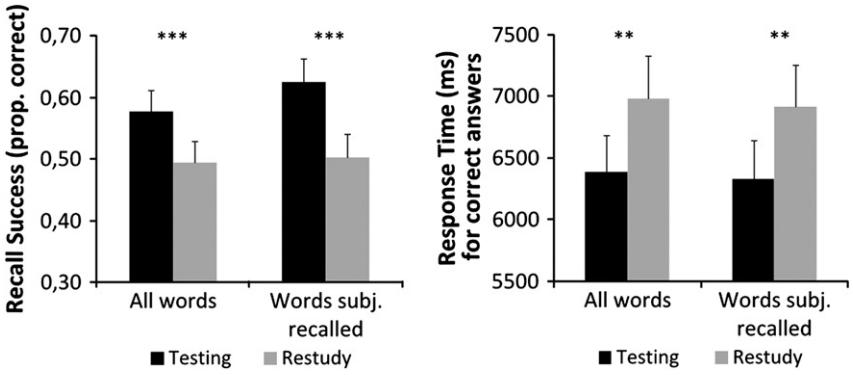

Fig. 2. Translation performance at the memory test seven days after testing and restudy practice. Proportion of words translated correctly and reaction times (for correct responses only) per practice condition, as measured on the final recall test after seven days. Results are displayed separately for all words (the two left bars of each figure) and for those words to which the participants responded "Yes, I know the translation" during practice (the two right bars of each figure). Error bars indicate standard errors of the mean. In all cases, performance was significantly better for the tested than for the restudied words. *** $p < .001$, ** $p < .01$.

that were predictive of later memory in the test condition but not in the restudy condition.

Discussion

In this study, we investigated neural correlates of testing effects by comparing testing and restudy practice in an fMRI experiment. Replicating previous behavioral results, delayed recall was better and faster for tested words than for restudied words (e.g., Karpicke and Roediger, 2008; Roediger and Karpicke, 2006a,b). Several areas in the brain were more active during testing than during restudy, including bilateral IFG and striatal areas. Areas that were more active during restudying than testing included the right MTG and bilateral IPL. Further analyses revealed that later memory was predicted by more activity in the left MTG and IPL during testing — but not restudying. Together, results show that testing improves memory retention more than restudying and that (1) this practice effect is related to greater activity in the IPL and MTG during testing but not during restudy, (2) that IFG activity is enhanced during testing in comparison to restudy, and that (3) increased activity in striatal and midbrain areas during testing may contribute to memory strengthening.

First, based on the notion that testing effects involve increased semantic elaboration of the connection between words and translations (Carpenter, 2009), we hypothesized that IPL and MTG would be more active during testing than restudying, and that activity in these areas would predict later memory. Results did not support the first prediction. On the contrary, activity in parts of IPL and MTG was higher during restudy than testing. However, the second prediction was partly confirmed: activity in the left IPL and MTG predicted later memory, yet only during testing and not during restudy. These results suggest that activity in the IPL and MTG reflects a cognitive function that is important for the beneficial effects of testing but not restudying.

IPL is an association cortex that is engaged in different higher cognitive functions, presumably supporting the integration of complex information and knowledge retrieval (Binder et al., 2009). During semantic elaboration, IPL is thought to integrate semantic information into context and to combine separate concepts into a larger coherent meaning (Lau et al., 2008). Memory studies have related IPL activity to both unsuccessful encoding and successful retrieval (e.g., Daselaar et al., 2009; Uncapher and Wagner, 2009). The relation with unsuccessful encoding has been attributed to increased elaboration of irrelevant information, such as during mind-wandering (Daselaar et al., 2009; Kim et al., 2010; Vannini et al., 2011). As a case in point, both the IPL and the MTG have been associated with the so-called default mode network (DMN), a set of brain areas that tends to be activated when thoughts are not focused on a specific task, for example, during rest and self-

Table 1
Brain regions showing more activity during testing than during restudy.

Cluster	Cluster size	*p*	Local maxima				
			Anatomical area	x	y	z	*t*
1	3382	<.0001	Left Anterior IFG, p. orbitalis (~BA 47)	−30	24	0	7.90
			Left Mid-IFG, p. triangularis (BA 45)	−40	24	24	5.48
			Left Posterior IFG, p. opercularis (BA 44)	−42	4	34	6.80
2	728	.0001	Right Anterior IFG p. orbitalis (~BA 47)	34	24	−4	7.38
			Right Mid-IFG, p. triangularis (BA 45)	42	20	10	3.47
3	1718	<.0001	Left supplementary motor area	−6	18	50	6.65
			Right middle cingulate	10	20	44	4.75
4	1764	<.0001	Right caudate nucleus	12	10	−2	5.89
			Left putamen	−12	6	−4	5.26
			Left thalamus	−6	−10	6	4.39
			Left midbrain	−8	−20	−12	4.07
			Right midbrain	8	−20	−12	4.06
5	970	<.0001	Left inferior parietal lobe	−32	−56	46	4.96
			Left middle occipital gyrus	−26	−72	42	4.66
6	1073	<.0001	Left lingual gyrus	−8	−82	10	4.48
			Right lingual gyrus	16	−68	8	3.86
			Right calcarine gyrus	12	−76	12	3.83

The table contains all clusters that were significantly more activated during testing than restudying, when combining later remembered and later forgotten trials over the three practice blocks. Statistical tests were performed with an uncorrected threshold of $p < .001$ at voxel-level, and a FWE-corrected threshold of $p < .05$ at cluster-level. The table lists the cluster-size in number of voxels, cluster-level p-value and information about local maxima (anatomical labels, MNI coordinates and t-values). BA = Brodmann area, IFG = inferior frontal gyrus.

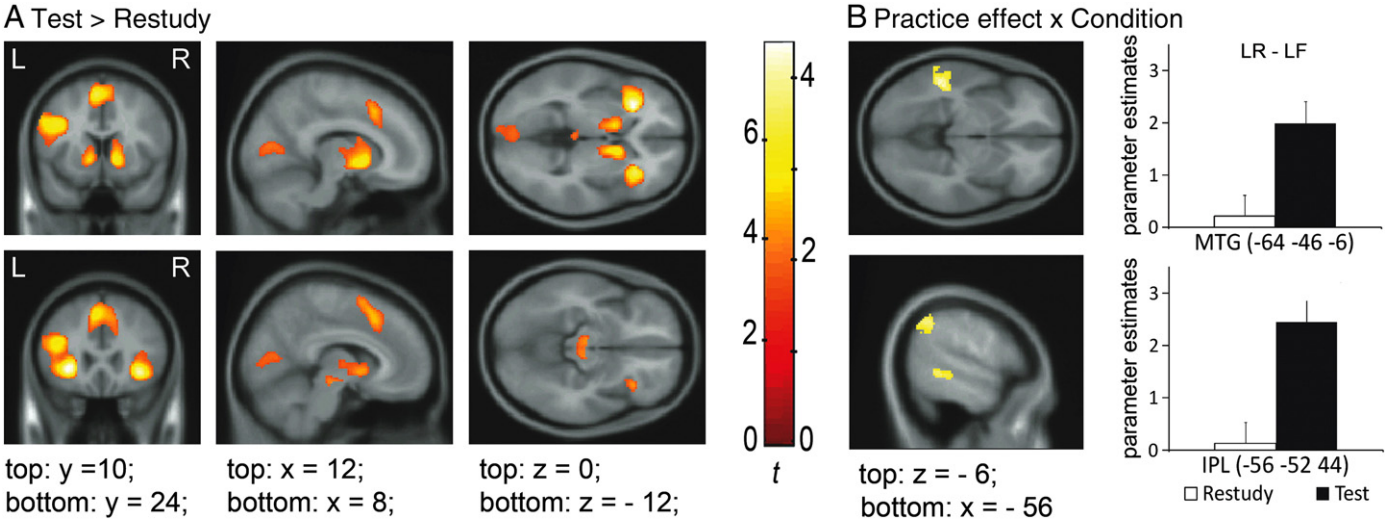

Fig. 3. Brain activity related to beneficial effects of testing. A. Clusters that were significantly more activated during testing than during restudying. Color coding as indicated on the left scale of the color map. B. Clusters in the left inferior parietal lobe (IPL) and middle temporal gyrus (MTG) that showed an interaction effect between practice condition and later memory. Activity in these regions during testing, but not during restudying was predictive of later memory. Color coding as indicated on the right scale of the color map. Statistical tests were performed with an uncorrected threshold of $p < .001$ at voxel-level, and corrected for multiple comparisons with a FWE-corrected threshold of $p < .05$ at cluster-level. Contrast estimates for the comparison of later remembered (LR) and later forgotten (LF) items are shown for two local maxima, error bars indicate 90% confidence intervals.

referential thoughts (e.g., Buckner et al., 2008; Mason et al., 2007). Other areas which were more activated during restudying than testing, such as the middle cingulate and medial orbitofrontal cortex, also show an overlap with the DMN, suggesting that some of the activation during restudying could reflect increased task-unrelated semantic processing.

On the other hand, there is an overlap between the DMN and cortical regions that are consistently engaged during successful episodic retrieval together with medial temporal lobe structures (Rugg and Vilberg, 2013). Areas of the DMN, including the angular gyrus in the IPL, tend to show greater activity during the recollection of stronger episodic memories than during familiarity responses to weaker memories (review in Kim, 2010). Moreover, the functional connectivity between DMN areas and the left hippocampus appears to increase during successful deep as compared to more shallow encoding, possibly reflecting the encoding of novel episodes into the larger scale self-referential DMN (Schott et al., 2013). Therefore, one interpretation of practice effects in IPL during testing could be the involvement of general recollection networks, an idea that is further supported by studies that link IPL activity to retrieval success: IPL activity increases when more information is retrieved (Vilberg and Rugg, 2008, 2009), feelings of remembering are strong (Wagner et al., 2005), or more attention is drawn towards retrieved information (Cabeza et al., 2008; Ciaramelli et al., 2010). So while overall higher activity in IPL during restudying than testing might reflect processes involved in self-referential thought or mind-wandering, the higher IPL activity during testing of later remembered than later forgotten words suggests that differences between retrieved representations predict later memory. Note that we only analyzed trials in which participants indicated that they successfully retrieved a translation. Activity is therefore likely to be driven by the amount or quality of the retrieved information and not by mere retrieval success.

The left MTG and neighboring regions are commonly associated with the long-term storage of lexical representations (Hagoort, 2005). Some argue that access to the meaning of words occurs in MTG (Jamal et al., 2012; Pugh et al., 2005), with this area acting as a store of conceptual features of semantic representations or as a hub that connects lexical representations to distributed semantic networks (Lau et al., 2008; Zhuang et al., 2011). Because activity in MTG predicted later memory only during testing and not during restudying, it seems that only processing of actively retrieved representations predicted later memory whereas processing of representations evoked by passive restudying did not. One possible explanation for this is that testing, more than

Table 2
Brain regions showing more activity during restudy than during testing.

Cluster	Cluster size	*p*	Local maxima				
			Anatomical area	x	y	z	*t*
1	1612	<.001	Right IPL	50	−70	28	4.97
			Right IPL	56	−56	46	3.93
			Right supramarginal gyrus (IPL; ~ BA 40)	54	−46	40	3.76
2	863	<.001	Right middle cingulate gyrus	8	−50	40	4.61
3	711	<.001	Right middle frontal gyrus	28	34	48	4.43
				26	28	42	4.26
4	1103	<.001	Left middle orbital gyrus	−8	58	4	4.39
			Right superior medial gyrus	12	60	10	4.36
			Right superior frontal gyrus	24	62	6	3.68
5	290	0.02	Right middle temporal gyrus	64	−16	−14	4.26
6	643	<.001	Left angular gyrus (IPL; ~ BA 39)	−54	−66	42	3.95
				−48	−60	36	3.88
				−58	−64	30	3.70

Structured like Table 1. BA = Brodmann area, IPL = Inferior parietal lobe.

Table 3
Brain regions showing (A) practice effect during restudy, i.e., more activity during restudying of words that were later remembered (LR) than during restudying of words that were later forgotten (LF); (B) practice effect during testing; (C) different practice effects during restudy and during testing.

Cluster	Cluster size	*p*	Local maxima				
			Anatomical area	x	y	z	*t*
(A) $LR_{RS} > LF_{RS}$							
1	242	0,0438	Left rectal gyrus	−2	40	−20	4.41
			Right rectal gyrus	6	52	−14	3.70
			Left superior orbital gyrus	−10	52	−14	3.36
(B) $LR_T > LF_T$							
1	5505	<.001	Left superior medial gyrus	−12	56	8	6.51
			Left superior frontal gyrus	−12	56	28	6.14
			Left superior frontal gyrus	−10	52	38	5.67
2	2353	<.001	Left middle temporal gyrus	−46	−56	28	6.06
			Left angular gyrus (IPL)	−54	−58	30	5.80
			Left Inferior Parietal Lobe	−56	−52	44	5.71
3	2926	<.001	Left middle cingulate cortex	−6	−52	40	6.02
			Left precuneus	−6	−56	22	4.85
4	1822	<.001	Right supramarginal gyrus	52	−50	40	5.66
			Right angular gyrus	44	−60	34	5.07
			Right superior temporal gyrus	56	−54	28	4.59
5	1405	<.001	Left middle temporal gyrus	−64	−46	−6	5.37
			Left middle temporal gyrus	−56	−42	−4	5.03
6	625	<.001	Right middle temporal gyrus	56	−14	−16	4.91
			Right inferior temporal gyrus	52	−26	−18	4.42
			Right inferior temporal gyrus	48	−10	−24	3.64
(C) $(LR_T–LF_T) > (LR_{RS}–LF_{RS})$							
1	253	0.04	Left middle temporal gyrus	−64	−46	−6	3.49
2	371	0.01	Left supramarginal gyrus (IPL)	−56	−52	44	3.93
			Left angular gyrus (IPL)	−54	−60	44	3.78
			Left supramarginal gyrus (IPL)	−56	−50	40	3.70
			Left angular gyrus (IPL)	−48	−68	44	3.67

Structured like Table 1. LR_T = activity during testing of later remembered items; LF_T = activity during testing of later forgotten items; LR_{RS} = activity during restudying of later remembered items; LF_{RS} = activity during restudying of later forgotten items. For all regions reported in the third section, the difference between later remembered and later forgotten items was larger in the testing than in the restudy condition. The reverse interaction (larger practice effect in restudy than in testing condition) showed no significant clusters. BA = Brodmann area. The supramarginal gyrus and the angular gyrus together form a part of the inferior parietal lobe (IPL).

restudying, activated relevant memory representations that facilitate later access to the translation, for example, mediators that link characteristics of the Swahili word to its translation (Carpenter, 2011; Pyc and Rawson, 2010).

In sum, activity in left IPL and MTG during testing but not during restudying was predictive of later memory. This does not support the idea that semantic processing is in general enhanced during testing in comparison to restudying, as was put forward in earlier testing effect papers (Carpenter and Delosh, 2006). Instead, results suggest that semantic processing during testing is more beneficial for memory than semantic processing during restudying, possibly because it is more focused on relevant associations. This explanation is in line with recent suggestions that testing improves later recall because it influences the specification of search sets that are activated in response to available retrieval cues, such that relevant target information is activated more effectively (Karpicke and Blunt, 2011; Karpicke and Smith, 2012; Karpicke and Zaromb, 2010). In terms of the present study, testing may have increased the suppression of incorrect translations that would otherwise be activated in response to the Swahili words and/or may have facilitated the activation of the correct translations. This idea that testing facilitated later recall by strengthening the association between the presented Swahili cues and the recalled translations is further supported by the behavioral outcome that tested words were translated significantly faster than restudied words on the final test.

A second major result that supports the conclusion that testing might selectively improve target associations was the enhanced activity in IFG during testing than restudying, which we had predicted based on accounts that mental effort is important for testing effects (e.g., Pyc and Rawson, 2009). IFG has repeatedly been related to intentional, non-automatic processing in memory studies (e.g., Race et al., 2009). During retrieval, IFG is thought to be involved in the controlled access to relevant information in memory and in the selection among competing representations (Badre and Wagner, 2007; Blumenfeld and Ranganath, 2007). Higher activation during testing than during restudying therefore supports the idea that testing involves more intentional, effortful processing than restudying. Possibly, the memory search during testing recruits control-processes in IFG for the activation of and selection among possible translations. Increased effort could also underlie the observed activations in lingual gyrus, which responds to visual processing demands (Mechelli et al., 2000) and in supplementary motor areas, which have been linked to effortful word selection processes in language production (Alario et al., 2006).

These results are particularly interesting in light of behavioral findings that testing effects increase with test difficulty: IFG activity during memory retrieval increases when cues are weak (e.g., Crescentini et al., 2010; Danker et al., 2008), and likewise, behavioral testing effects increase when cues are weak (Carpenter, 2009; Carpenter and Delosh, 2006). Vice versa, IFG activity decreases during repeated retrieval acts (Petersson et al., 1999), and likewise, the amount of memory improvement per retrieval act decreases with repetition, especially when the delay between retrievals is short (Pyc and Rawson, 2009). These neural and behavioral results have both been explained with changing demands on controlled, effortful processing (e.g., Danker et al., 2008; Kelly and Garavan, 2005; Pyc and Rawson, 2009). Interpreting IFG activity in the present study in terms of enhanced cognitive control is thus in line with previous imaging and behavioral studies about repeated testing practice as well as with theoretical claims that testing constitutes a desirable difficulty during learning that improves memory (Bjork and Bjork, 1992).

However, whereas activity in IFG during encoding has consistently been related to effective memory formation, in particular for verbal information (meta-analysis by Kim, 2011), we found only indirect proof of such a relation in this study: IFG activity was higher and later memory was better for tested than for restudied items but there was no direct relation between IFG activity during practice and later memory (i.e., no practice effect). This could be due to the fact that – unlike previous studies – we measured brain activity during additional practice of stimuli that had already been studied extensively before. It is plausible that learners invested more effort to practice words that they found difficult to remember than to practice words that they found easy, which could conceal positive effects of effort on memory if the difficult words were more likely to be forgotten.

In sum, testing increased activity in IFG in comparison to restudying, possibly reflecting higher demands on effortful control processes necessary for the selective activation of the correct translations, but the amount of this processing as such was not predictive of better memory retention.

Additional regions that were involved in testing more than restudying included parts of the midbrain and the ventral striatum. This is interesting because these are key structures of the brain's motivation and reward-system (Shohamy and Adcock, 2010). Dopaminergic neurons that project from tegmental areas in the midbrain to the ventral striatum highlight motivationally significant information (Camara et al., 2009), and direct attention toward relevant or 'adaptive' information during memory encoding (Wittmann et al., 2008; for a review, see Shohamy and Adcock, 2010; Wittmann et al., 2005). Increased activity in these areas could reflect an additional mechanism by which testing strengthens the memory trace by highlighting information as relevant and enhancing attention. This is in line with speculations that during testing, interactions between the hippocampus and dopaminergic neurons in ventral tegmental midbrain areas could enhance long-term potentiation in the hippocampus and thereby learning (Roediger and Butler, 2011). In addition, genetic determinants of dopamine projections to the prefrontal cortex have been related to retrieval-induced suppression of irrelevant information, which presumably reduces future interference (Wimber et al., 2011). As dopaminergic activations are higher during more effortful tasks, it has been speculated that dopaminergic regions might be involved in a gating mechanism that adjusts the amount of cognitive resources for the processing of incoming information (Boehler et al., 2011). Involvement of such a gating mechanism would offer a plausible explanation for testing effects from an evolutionary point of view: information that is readily available in the environment (as during restudying), is likely to remain available in the future. In contrast, information that must be retrieved from memory with effort (as during testing) is likely to cost cognitive capacities again during future retrievals. Therefore, investing resources to better remember tested information is more useful on average than to remember restudied information, because remembering tested information is more likely to reduce future processing costs.

Conclusion

We report three major findings on mechanisms potentially underlying testing effects: first, semantic association areas in the left IPL and MTG were more active during testing of later remembered than later forgotten words, but showed no such relation to later memory for the restudied items. Activity in these areas might reflect the selective enrichment of semantic associations that improve later access to the target-information during testing. Second, testing increased activity in IFG in comparison to restudying. This supports claims that testing requires more effortful cognitive control than restudying due to the suppression of irrelevant responses and the selective activation of target information. Third, areas in the ventral striatum and midbrain were more active during testing than during restudying, which could reflect activity that supports prefrontal selection processes during memory retrieval as well as motivation and reward circuits that strengthen memory retention. To conclude, the present study improves insight into the neural correlates of testing effects; it thereby adds to explanations of behaviorally established testing effects and further encourages the use of tests in educational practice.

Acknowledgments

This research was supported by a grant from the National Initiative Brain & Cognition, Netherlands Organization for Scientific Research (NWO grant number 056-33-014). The authors thank Paul Gaalman for his technical support.

The authors have declared that no competing interests exist.

References

Alario, F.X., Chainay, H., Lehericy, S., Cohen, L., 2006. The role of the supplementary motor area (SMA) in word production. Brain Res. 1076, 129–143.
Badre, D., Wagner, A.D., 2007. Left ventrolateral prefrontal cortex and the cognitive control of memory. Neuropsychologia 45, 2883–2901.
Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb. Cortex 19, 2767–2796.
Bjork, R.A., Bjork, E.L., 1992. A new theory of disuse and an old theory of stimulus fluctuation. In: Healy, A., Kosslyn, S., Shiffrin, R. (Eds.), From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes. Erlbaum, Hillsdale, NJ, pp. 35–67.
Blumenfeld, R.S., Ranganath, C., 2007. Prefrontal cortex and long-term memory encoding: an integrative review of findings from neuropsychology and neuroimaging. Neuroscientist 13, 280–291.
Boehler, C.N., Hopf, J.-M., Krebs, R.M., Stoppel, C.M., Schoenfeld, M.A., Heinze, H.-J., Noesselt, T., 2011. Task-load-dependent activation of dopaminergic midbrain areas in the absence of reward. J. Neurosci. 31, 4955–4961.
Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L., 2008. The brain's default network: anatomy, function, and relevance to disease. ANYAS 1124, 1–38.
Cabeza, R., Ciaramelli, E., Olson, I.R., Moscovitch, M., 2008. The parietal cortex and episodic memory: an attentional account. Nat. Rev. Neurosci. 9, 613–625.
Camara, E., Rodriguez-Fornells, A., Ye, Z., Münte, T.F., 2009. Reward networks in the brain as captured by connectivity measures. Front. Neurosci. 3, 350–362.
Carpenter, S.K., 2009. Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. J. Exp. Psychol. Learn. Mem. Cogn. 35, 1563–1569.
Carpenter, S.K., 2011. Semantic information activated during retrieval contributes to later retention: support for the mediator effectiveness hypothesis of the testing effect. J. Exp. Psychol. Learn. Mem. Cogn. 37, 1547–1552.
Carpenter, S.K., Delosh, E.L., 2006. Impoverished cue support enhances subsequent retention: support for the elaborative retrieval explanation of the testing effect. Mem. Cognit. 34, 268–276.
Ciaramelli, E., Grady, C., Levine, B., Ween, J., Moscovitch, M., 2010. Top-down and bottom-up attention to memory are dissociated in posterior parietal cortex: neuroimaging and neuropsychological evidence. J. Neurosci. 30, 4943–4956.
Crescentini, C., Shallice, T., Macaluso, E., 2010. Item retrieval and competition in noun and verb generation: an fMRI study. J. Cogn. Neurosci. 22, 1140–1157.
Danker, J.F., Gunn, P., Anderson, J.R., 2008. A rational account of memory predicts left prefrontal activation during controlled retrieval. Cereb. Cortex 18, 2674–2685.
Daselaar, S.M., Prince, S.E., Dennis, N.A., Hayes, S.M., Kim, H., Cabeza, R., 2009. Posterior midline and ventral parietal activity is associated with retrieval success and encoding failure. Front. Hum. Neurosci. 3, 350–362.
Eriksson, J., Kalpouzos, G., Nyberg, L., 2011. Rewiring the brain with repeated retrieval: a parametric fMRI study of the testing effect. Neurosci. Lett. 505, 36–40.
Gabrieli, J.D.E., Desmond, J.E., Demb, J.B., Wagner, A.D., Stone, M.V., Vaidya, C.J., Glover, G.H., 1996. Functional magnetic resonance imaging of semantic memory processes in the frontal lobes. Psychol. Sci. 7, 278–283.
Hagoort, P., 2005. On Broca, brain, and binding: a new framework. Trends Cogn. Sci. 9, 416–423.
Hashimoto, T., Usui, N., Taira, M., Kojima, S., 2011. Neural enhancement and attenuation induced by repetitive recall. Neurobiol. Learn. Mem. 96, 143–149.
Hayasaka, S., Nichols, T.E., 2003. Validating cluster size inference: random field and permutation methods. NeuroImage 20, 2343–2356.
Jamal, N.I., Piche, A.W., Napoliello, E.M., Perfetti, C.A., Eden, G.F., 2012. Neural basis of single-word reading in Spanish–English bilinguals. Hum. Brain Mapp. 33, 235–245.
Karpicke, J.D., Blunt, J.R., 2011. Retrieval practice produces more learning than elaborative studying with concept mapping. Science 331, 772–775.
Karpicke, J.D., Roediger III, H.L., 2008. The critical importance of retrieval for learning. Science 319, 966–968.
Karpicke, J.D., Smith, M.A., 2012. Separate mnemonic effects of retrieval practice and elaborative encoding. J. Mem. Lang. 67, 17–29.
Karpicke, J.D., Zaromb, F.M., 2010. Retrieval mode distinguishes the testing effect from the generation effect. J. Mem. Lang. 62, 227–239.
Kelly, A., Garavan, H., 2005. Human functional neuroimaging of brain changes associated with practice. Cereb. Cortex 15, 1089.

Kim, H., 2010. Dissociating the roles of the default-mode, dorsal, and ventral networks in episodic memory retrieval. NeuroImage 50, 1648–1657.
Kim, A.S.N., 2011. Neural activity that predicts subsequent memory and forgetting: a meta-analysis of 74 fMRI studies. NeuroImage 54, 2446–2461.
Kim, A.S.N., Daselaar, S.M., Cabeza, R., 2010. Overlapping brain activity between episodic memory encoding and retrieval: roles of the task-positive and task-negative networks. NeuroImage 49, 1045–1054.
Lau, E.F., Phillips, C., Poeppel, D., 2008. A cortical network for semantics:(de) constructing the N400. Nat. Rev. Neurosci. 9, 920–933.
Mason, M.F., Norton, M.I., Van Horn, J.D., Wegner, D.M., Grafton, S.T., Macrae, C.N., 2007. Wandering minds: the default network and stimulus-independent thought. Science 315, 393–395.
Mechelli, A., Humphreys, G.W., Mayall, K., Olson, A., Price, C.J., 2000. Differential effects of word length and visual contrast in the fusiform and lingual gyri during reading. Proc. R. Soc. Lond. B Biol. Sci. 267, 1909–1913.
Petersson, K.M., Elfgren, C., Ingvar, M., 1999. Dynamic changes in the functional anatomy of the human brain during recall of abstract designs related to practice. Neuropsychologia 37, 567–587.
Pugh, K.R., Sandak, R., Frost, S.J., Moore, D., Mencl, W.E., 2005. Examining reading development and reading disability in English language learners: potential contributions from functional neuroimaging. Learn. Disabil. Res. Pract. 20, 24–30.
Pyc, M.A., Rawson, K.A., 2009. Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? J. Mem. Lang. 60, 437–447.
Pyc, M.A., Rawson, K.A., 2010. Why testing improves memory: mediator effectiveness hypothesis. Science 330, 335.
Race, E.A., Kuhl, B.A., Badre, D., Wagner, A.D., 2009. The dynamic interplay between cognitive control and memory. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences. MIT Press, Cambridge, MA, pp. 705–724.
Roediger, H.L., Butler, A.C., 2011. The critical role of retrieval practice in long-term retention. Trends Cogn. Sci. 15, 20–27.
Roediger, H.L., Karpicke, J.D., 2006a. The power of testing memory: basic research and implications for educational practice. Perspect. Psychol. Sci. 1, 181–210.
Roediger, H.L., Karpicke, J.D., 2006b. Test-enhanced learning: taking memory tests improves long-term memory. Psychol. Sci. 17, 249–255.
Rugg, M.D., Vilberg, K.L., 2013. Brain networks underlying episodic memory retrieval. Curr. Opin. Neurobiol. 23, 255–260.
Schott, B.H., Wüstenberg, T., Wimber, M., Fenker, D.B., Zierhut, K.C., Seidenbecher, C.I., Heinze, H.-J., Walter, H., Düzel, E., Richardson-Klavehn, A., 2013. The relationship between level of processing and hippocampal–cortical functional connectivity during episodic memory formation in humans. Hum. Brain Mapp. 34, 407–424.
Shohamy, D., Adcock, R.A., 2010. Dopamine and adaptive memory. Trends Cogn. Sci. 14, 464–472.
Thomas, R.C., McDaniel, M.A., 2013. Testing and feedback effects on front-end control over later retrieval. J. Exp. Psychol. Learn. Mem. Cogn. 39, 437–450.
Uncapher, M.R., Wagner, A.D., 2009. Posterior parietal cortex and episodic encoding: insights from fMRI subsequent memory effects and dual-attention theory. Neurobiol. Learn. Mem. 91, 139–154.
Vannini, P., O'Brien, J., O'Keefe, K., Pihlajamäki, M., LaViolette, P., Sperling, R., 2011. What goes down must come up: role of the posteromedial cortices in encoding and retrieval. Cereb. Cortex 21, 22–34.
Vilberg, K.L., Rugg, M.D., 2008. Memory retrieval and the parietal cortex: a review of evidence from a dual-process perspective. Neuropsychologia 46, 1787–1799.
Vilberg, K.L., Rugg, M.D., 2009. Left parietal cortex is modulated by amount of recollected verbal information. Neuroreport 20, 1295–1299.
Wagner, A.D., Schacter, D.L., Rotte, M., Koutstaal, W., Maril, A., Dale, A.M., Rosen, B.R., Buckner, R.L., 1998. Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. Science 281, 1188–1191.
Wagner, A.D., Shannon, B.J., Kahn, I., Buckner, R.L., 2005. Parietal lobe contributions to episodic memory retrieval. Trends Cogn. Sci. 9, 445–453.
Whitney, C., Kirk, M., O'Sullivan, J., Lambon Ralph, M.A., Jefferies, E., 2011. The neural organization of semantic control: TMS evidence for a distributed network in left inferior frontal and posterior middle temporal gyrus. Cereb. Cortex 21, 1066–1075.
Wimber, M., Schott, B.H., Wendler, F., Seidenbecher, C.I., Behnisch, G., Macharadze, T., Bauml, K.H.T., Richardson-Klavehn, A., 2011. Prefrontal dopamine and the dynamic control of human long-term memory. Transl. Psychiatry 1, e15.
Wittmann, B.C., Schott, B.H., Guderian, S., Frey, J.U., Heinze, H.J., Düzel, E., 2005. Reward-related FMRI activation of dopaminergic midbrain is associated with enhanced hippocampus-dependent long-term memory formation. Neuron 45, 459–467.
Wittmann, B.C., Schiltz, K., Boehler, C.N., Düzel, E., 2008. Mesolimbic interaction of emotional valence and reward improves memory formation. Neuropsychologia 46, 1000–1008.
Zhuang, J., Randall, B., Stamatakis, E.A., Marslen-Wilson, W.D., Tyler, L.K., 2011. The interaction of lexical semantics and cohort competition in spoken word recognition: an fMRI study. J. Cogn. Neurosci. 23, 3778–3790.